Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W435–W439***
doi:10.1093/nar/gkl200

AUGUSTUS: *ab initio* prediction of alternative transcripts

Mario Stanke*, Oliver Keller[1], Irfan Gunduz[2], Alec Hayes[2], Stephan Waack[1] and Burkhard Morgenstern

Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik, Goldschmidtstrasse 1, 37077 Göttingen, Germany, [1]Institut für Informatik, Lotzestrasse 16-18, 37083 Göttingen, Germany and [2]Philip Morris USA, Research Center, Richmond, VA 23261, USA

Received February 14, 2006; Revised and Accepted March 21, 2006

ABSTRACT

AUGUSTUS is a software tool for gene prediction in eukaryotes based on a Generalized Hidden Markov Model, a probabilistic model of a sequence and its gene structure. Like most existing gene finders, the first version of AUGUSTUS returned one transcript per predicted gene and ignored the phenomenon of alternative splicing. Herein, we present a WWW server for an extended version of AUGUSTUS that is able to predict multiple splice variants. To our knowledge, this is the first *ab initio* gene finder that can predict multiple transcripts. In addition, we offer a motif searching facility, where user-defined regular expressions can be searched against putative proteins encoded by the predicted genes. The AUGUSTUS web interface and the downloadable open-source stand-alone program are freely available from http://augustus.gobics.de.

INTRODUCTION

Despite considerable efforts in the Bioinformatics community, the performance of existing gene prediction tools is still not satisfactory. In current genome projects, a common approach for gene finding is the following. Several sets of gene predictions are compiled, usually from different gene finders trained specifically for the species at hand. Further, alignments from ESTs and proteins to the genome are constructed. Finally, the predictions and the alignments are combined to find plausible gene structures, either manually or by using meta tools that combine several predictions and alignments [e.g. (1)].

AUGUSTUS is a gene finder based on a Generalized Hidden Markov Model (GHMM) (2,3). The original version of the program was a purely *ab initio* method, i.e. its prediction was based on information contained in the genomic sequence to be analyzed. An extended version of the program is able to use additional extrinsic information, for example matches to protein databases or alignments of genomic sequences, to improve the prediction accuracy (4). At the recent *EGASP* workshop in Cambridge, UK, a systematic evaluation of existing gene finders for the human genome has been carried out based on a large set of well-annotated parts of the human genome (5). At this workshop, AUGUSTUS turned out to be the best program in the category of *ab initio* gene prediction. Its performance could be further improved by using BLAST (6) hits to EST or protein sequences and alignments of syntenic genomic sequences using DIALIGN (7,8); in this category, however, the program was outperformed by *N-Scan* (9), a new program based on multiple alignments of genomic sequences. Compared to more traditional approaches, gene-finding methods based on genomic sequence alignments have a considerable advantage since they do not depend on EST or protein sequences or statistical models of gene structures (10–13). On the other hand, alignment-based methods work only if genome sequences at an appropriate evolutionary distance are available. Although the performance of *ab initio* gene-prediction methods is usually improved if information from comparative sequence analysis is added, *ab initio* gene prediction remains highly important since for many newly sequenced genomes, few EST or related genomic sequences sequences are available and comparison to protein sequences can find only those genes that have close relatives in existing databases.

To make AUGUSTUS available to the research community, we set up a WWW server at Göttingen Bioinformatics Compute Server (GOBICS) (14,15). Like most gene-prediction methods that are currently available, earlier versions of AUGUSTUS predicted exactly one transcript per gene and ignored the fact that one gene often yields more than one distinct mRNA product. It has been estimated that 40–60% of all human genes have alternative splice forms. Of those genes 70–88% of alternative splices change the protein product; the remaining splice variants differ in the untranslated regions only (16). Thus, it is important to have gene-finding tools that are able to deal with this phenomenon. The program SLAM (17), for example, predicts

*To whom correspondence should be addressed. Tel: +1 831 459 5232; Fax: +1 832 459 1809; Email: mstanke@gwdg.de

alternative splice variants. This program, however, is based on alignments of genomic sequences, and it requires two syntenic genomic sequences as input data. We recently installed a new version of AUGUSTUS at our server that can predict multiple transcripts for predicted genes. To our knowledge, this is the first *ab initio* gene finder that can predict multiple transcripts, and our web server is the only gene prediction web server with this option.

With our new alternative-transcripts option, the user can control the number of predicted splice variants per gene. This way, it is possible to influence sensitivity and specificity of the program output. If predicted genes or transcripts are automatically evaluated and post-processed, high prediction sensitivity may be desirable to increase the number of candidate genes that are to be analyzed, even if this increases the number of false-positive predictions. In contrast, if expensive experiments are carried out based on computationally predicted genes, it is preferable to have highly specific tools that minimize the risk of false-positive predictions. Thus, a good gene-finding method should allow the user to chose between high sensitivity and high specificity. At our server, this can be done by specifying the maximum number of predicted splice variants. In addition, we implemented a motif-searching option at our server where predicted genes can be searched for user-specified regular expressions, e.g. PROSITE patterns.

MATERIALS AND METHODS

Sampling and posterior probabilities

Let the term parse denote a segmentation of the input DNA sequence s into exons, introns and intergenic regions. Without considering alternative transcripts, each parse would define a distinct gene structure on s. For each species for which AUGUSTUS has been trained, AUGUSTUS has one GHMM. As usual in HMM-based gene prediction, this model defines a probability distribution over all pairs of a DNA sequence s and a parse φ. Now, let s be a given fixed input DNA sequence s in which we want to find gene structures. The model implicitly defines a conditional probability distribution over all parses. Let

$$p(\varphi \mid s) \qquad \mathbf{1}$$

denote the probability of parse φ given input sequence s. We call $p(\varphi \mid s)$ the posterior probability of φ. If the probabilistic modelling is good, the posterior probability is relatively high for true or reasonable parses φ. We call the parse with the highest posterior probability the Viterbi parse according to the name of the algorithm that is used to find it (18). For an input sequence s, most HMM-based gene prediction programs output one single parse, namely the gene structure corresponding to the Viterbi parse.

Random sampling has been used in the context of gene prediction, e.g. by Alexanderson *et al*. (19). To produce alternative splice variants for a sequence s, AUGUSTUS randomly samples n parses $\varphi_1, \varphi_2, \ldots, \varphi_n$ using a sampling algorithm that has been described in (3). In each of these n random experiments, a parse φ is picked with probability $p(\varphi \mid s)$. Usually, many sampled parses share exons, introns or transcripts. The sampled parses are used to estimate posterior probabilities of exons, introns, transcripts and genes. The posterior probobability of an exon or intron is the probability of this exon or intron to be part of the sampled parse with exactly the same boundaries and on the same strand. The posterior probability of a transcript is the probability with which the sampled parse contains a transcript that is from the start codon to the stop codon completely identical to this transcript. With the posterior probability of a gene we denote the probability that the sampled parse contains some coding region on the strand of the gene within the boundaries of the gene. While the posterior probabilities of exons, introns and transcripts could be computed theoretically using the backward and forward algorithm, the probability of a gene cannot be easily computed. For consistent results, we therefore estimated all posterior probabilities using our sampling approach. In our approach, posterior probabilities of exons, introns, transcripts and genes are estimated by their relative frequencies in the sampling procedure. For example, if an exon occurres 80 times in $n = 100$ sampled parses, then the posterior probability is estimated as 80%.

Alternative transcripts

Our aim was to construct a predicted set of genes with likely alternative transcripts (Figure 1). The number of alternative transcripts for a gene should be different from gene to gene, depending on how many likely alternatives exist. We first compile a set of transcripts by taking all transcripts from the Viterbi parse ('Viterbi transcripts') and from all sampled parses. Then we estimate the posterior probabilities of all transcripts, exons, introns and as described above. We discard transcripts where the coding sequence has a length smaller than a certain minimum length $L_{\min}$ where the default value for $L_{\min}$ is 102 bp. Further, we apply the following filtering criteria to the non-Viterbi transcripts to retain only the likely alternatives.

- We throw away transcripts where any exon or intron has a posterior probability below some constant $P_{\min}$.
- We throw away transcripts where the geometric mean of the posterior probability of all exons and introns is below a constant $P^{av}_{\min}$
- If more than $T_{\max}$ transcripts overlap at the same position, we keep only $T_{\max}$ of them, giving highest priority to Viterbi-transcripts and further sorting by mean posterior probability of exons and introns.

The third criterion ensures that only $T_{\max}$ tracks are needed when the predictions are displayed in a genome browser (in Figure 2, $T_{\max} = 3$). Finally, predicted transcripts are clustered to genes in such a way that any two overlapping transcripts on the same strand are in the same gene. Our web server has four different options for the number of reported alternative transcripts, single transcript and few, medium number or many transcripts. For these options, the following parameter settings are used:

- single transcript: only Viterbi-transcripts are reported.
- few transcripts: $P_{\min} = 20\%$, $P^{av}_{\min} = 50\%$ $T_{\max} = 2$
- medium number of transcripts: $P_{\min} = 8\%$ $P^{av}_{\min} = 40\%$, $T_{\max} = 3$
- many transcripts: $P_{\min} = 8\%$, $P^{av}_{\min} = 30\%$, $T_{\max} = 20$

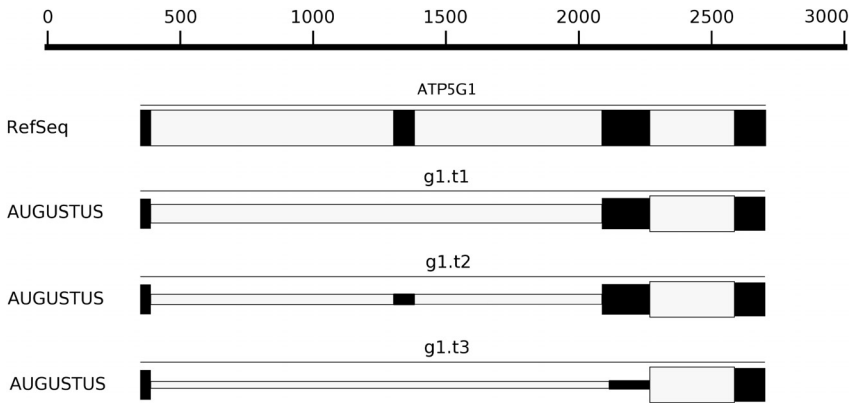

Figure 1. The human gene *ATP5G1* and the AUGUSTUS *ab initio* prediction for this region. The first transcript (g1.t1) is also the one predicted by standard AUGUSTUS using the Viterbi algorithm only. It misses the second exon of the gene. The second transcript (g1.t2) contains that exon and is correct. The height of a box (black: exon, light gray: intron) reflects the posterior probability of that exon or intron: The higher the posterior probability, the higher the box.

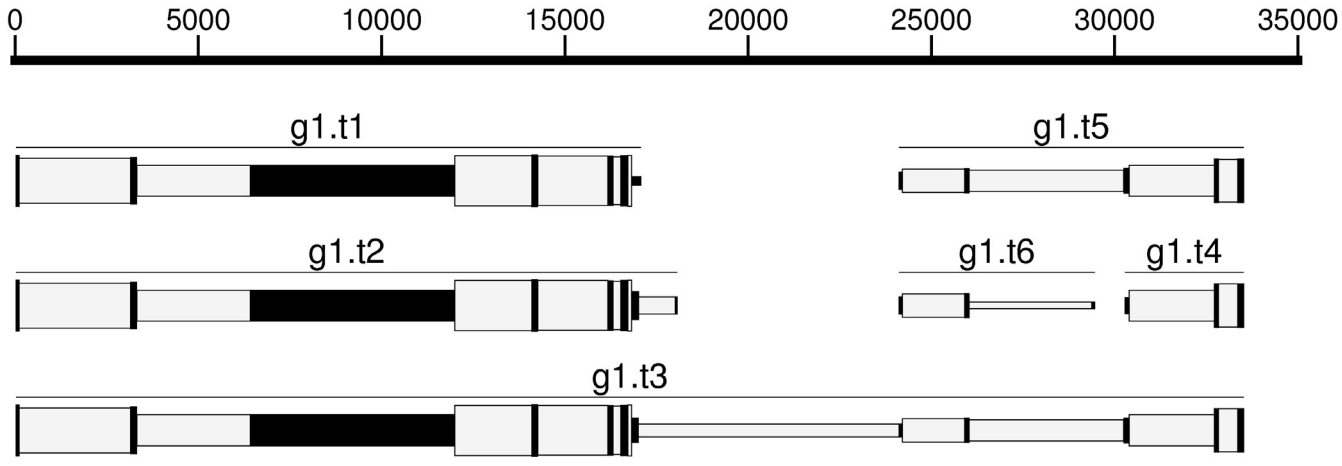

Figure 2. Region of a human gene on the forward strand for which AUGUSTUS predicted six transcripts (gene SON, chromosome 21, 33 837 000–33 872 000, ncbi build 35). The long intron of transcript g1.t3 containing position 20 000 has low posterior probability. Thus, the model is unsure whether this is actually one gene, two or three genes. In fact, for this gene there exists EST evidence both for the short transcript g1.t1 and for longer transcripts with exons mostly agreeing with those predicted above.

WEB SERVER DESCRIPTION

Input

At the AUGUSTUS web server, the user can upload their sequences in FASTA format or paste them into a web form. The maximal total length of the sequences submitted to the server is 3 million base pairs. AUGUSTUS has species-specific parameter sets that can be chosen at the web site. For the following species, pre-calculated sets of parameters are available: *Homo sapiens, Drosophila melanogaster, Arabidopsis thaliana, Brugia malayi, Coprinus cinereus, Tribolium castaneum, Schistosoma mansoni, Tetrahymena thermophila* and *Galdieria sulphuraria*.

Output

AUGUSTUS outputs its results both in graphical and text format. The text output consists of exon, intron, transcript and gene boundaries in the common General Feature Format (GFF) as well as predicted amino acid sequences and predicted coding sequences in FASTA format. The results page of our web server shows for each sequence a thumbnail and links to images in pdf and postscript format similar to the example shown in Figures 1 and 2. The graphical output is generated with the program gff2ps (20) from the text output.

SEARCHING FOR REGULAR EXPRESSIONS

To analyze putative protein products of predicted genes, our server offers a pattern-searching option. Here, the user can enter arbitrary patterns as regular expressions; these patterns are then searched against all predicted splice variants from all genes predicted in the input sequence. This can be helpful if the user is looking for members of a particular protein family with conserved positions that are already described by a pattern, for example as an entry in the PROSITE

Table 1. Percentage of correctly predicted human exons and introns in coding regions grouped by their posterior probability

Posterior probability P	Exon specificity
$0 \leqslant P \leqslant 50\%$	$46/242 \approx 19.0\%$
$50\% < P \leqslant 70\%$	$132/356 \approx 37.1\%$
$70\% < P \leqslant 80\%$	$84/175 \approx 48.0\%$
$80\% < P \leqslant 90\%$	$171/275 \approx 62.2\%$
$90\% < P \leqslant 95\%$	$140/195 \approx 71.8\%$
$95\% < P \leqslant 99\%$	$338/422 \approx 80.1\%$
$99\% < P \leqslant 100\%$	$545/612 \approx 89.1\%$
Total	$1456/2277 \approx 63.9\%$

For example, out of 2277 exons predicted by AUGUSTUS, 422 had a posterior probability between 0.95 and 0.99 of which 338 (80.1%) matched exactly an annotated exon. Here, AUGUSTUS was set to predict only one transcript per gene (no alternatives). As reference annotation the ENCODE test set with 296 genes and 649 transcripts was used, which is a challenging test set: the exon-level specificities of AUGUSTUS, GENEID, GENEZILLA and GENSCAN are 63.9, 61.1, 50.3 and 46.4%, respectively.

Table 2. Accuracy values of variants of AUGUSTUS on the ENCODE test set with 296 genes and an average of 2.2 transcripts per gene

	Single transcript	Few transcripts	Medium number of transcripts	Many transcripts
Gene sensitivity	0.233	0.273	0.294	0.345
Gene specificity	0.170	0.200	0.211	0.239
Transcript sensitivity	0.106	0.125	0.137	0.165
Transcript specificity	0.170	0.144	0.112	0.053
Exon sensitivity	0.527	0.540	0.557	0.600
Exon specificity	0.639	0.615	0.571	0.490
Base sensitivity	0.775	0.779	0.790	0.814
Base specificity	0.764	0.757	0.738	0.705
Average transcripts/gene	1.0	1.4	2.0	5.1

We used AUGUSTUS with the original single-transcript option and with the new options for 'few transcripts', 'medium number of transcripts' and 'many transcripts'.

database. In particular, in connection with the above described option for alternative splice forms, our pattern-searching option can be applied to filter out those splice variants that are consistent with the described pattern.

Following the PROSITE syntax, the conserved positions of a protein family may be specified as in the following example describing the hydrophobin family: <{C}(17,85)-C-{C}(5,10)-C-C-{C}(11,44)-C-{C}(8,23)-C-{C}(5,9)-C-C-{C}(6,18)-C-{C}(2,13)>. This pattern consists of eight conserved cystein residues with certain restrictions on the space between two subsequent cysteins. If a genome is to be searched for members of a protein family characterized by a regular expression, it is advisable to let AUGUSTUS predict a large number of splice variants to increase the prediction sensitivity, possibly at the expense of its specificity. With this strategy, the number of false positive splice variants may be increased, but a large number of predicted transcripts also increases the probability of finding all instances of the specified motif in the genome sequence under study.

RESULTS

We tested our method on a large set of test data from the recent EGASP workshop that was organized in the context of the ENCODE project (5). This dataset comprises a total of 296 genes with an average of 2.2 transcripts per gene. Table 1 shows the relation between the posterior probability of exons as defined above and the specificity at the exon level. As one may expect, the posterior probability is a good indicator of exon specificity. Predicted exons with low posterior probability are much less likely to be correct than predicted exons with high posterior probability. Therefore, the posterior probability gives a good criterion for prioritising putative exons, e.g. for experimental verification.

Table 2 summarizes the prediction accuracy of AUGUSTUS on the EGASP test data. We used AUGUSTUS with the original single-transcript option and with the new options for 'few transcripts', 'medium number of transcripts' and 'many transcripts'. Sensitivity and specificity of these program versions are given at four different levels, namely at the base, exon, transcript and gene level. At the gene level, a gene is considered a true positive if one predicted transcript coincides with an existing transcript of this gene. As expected, the program sensitivity increases at all four levels if the number of transcripts per gene is increased. Correspondingly, the specificity at the base, exon and transcript level decreases if the number of predicted transcript increases. In contrast, at the gene level the specificity of AUGUSTUS increases if more transcripts per gene are predicted. The reason for this seemingly paradoxical result is as follows: with an increased number of predicted transcripts, it is more likely that one of the real transcripts in a gene is matched by a predicted transcript. The number of predicted genes, on the other hand, does not increase significantly if more transcripts are predicted because most of the additionally predicted transcripts belong to one of the previously predicted genes. Thus, the ratio of true positive genes to predicted genes increases if the number of predicted transcripts is increased.

Letting AUGUSTUS predict many transcripts is most useful when the focus is on finding at least one correct splice form for a gene: as shown in Table 2, the gene-level sensitivity increases from 23.3 to 34.5% if the option 'many transcripts' is used (on average, 5.1 transcripts per gene are predicted with this option). This is substantially more than the gene-level sensitivity of standard gene finders; e.g. for the commonly used programs GENSCAN (21), GENEID (22) and GENEZILLA (23), the gene-level sensitivity values on the EGASP dataset are 15.5, 10.5 and 19.6%, respectively.

ACKNOWLEDGEMENTS

The option of predicting multiple transcripts per gene was developed in part when training AUGUSTUS for *Nicotiana tabacum* in the context of the Tobacco Genome Initiative http://tgi.ncsu.edu, (24) in Philip Morris USA. This work was supported by BMBF grant 01AK803G (Medigrid) to B.M. and by a fellowship within the Postdoc-Program of the German Academic Exchange Service (DAAD) to M.S. Funding to pay the Open Access publication charges for this article was provided by Philip Morris USA.

Conflict of interest statement. None declared.

REFERENCES

1. Allen,J.E., Pertea,M. and Salzberg,S.L. (2004) Computational gene prediction using multiple sources of evidence. *Genome Res*., **14**, 142–148.
2. Stanke,M. and Waack,S. (2003) Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics (ECCB 2003 special issue)*, **19**, ii215–ii225.
3. Stanke,M. (2004) Gene Prediction with a Hidden Markov Model. PhD Thesis. Universität Göttingen.
4. Stanke,M., Schöffmann,O., Morgenstern,B. and Waack,S. (2006) Gene prediction in eukaryotes with a Generalized Hidden Markov Model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.
5. Guigó,R. and Reese,M.G. (2005) EGASP: collaboration through competition to find human genes. *Nature Meth*., **2**, 575–577.
6. Altschul,S.F., Gish,W., Miller,W., Myers,E.M. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
7. Brudno,M., Chapman,M., Göttgens,B., Batzoglou,S. and Morgenstern,B. (2003) Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, **4**, 66.
8. Morgenstern,B. (2004) DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. *Nucleic Acids Res*., **32**, W33–W36.
9. Gross,S.S. and Brent,M.R. (2005) Using Multiple Alignments to Improve Gene Prediction. *Proceedings RECOMB'05*, pp. 374–388.
10. Batzoglou,S., Pachter,L., Mesirov,J.P., Berger,B. and Lander,E.S. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res*., **10**, 950–958.
11. Bafna,V. and Huson,D.H. (2000) The conserved exon method for gene finding. *Bioinformatics*, **16**, 190–202.
12. Korf,I., Flicek,P., Duan,D. and Brent,M.R. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, S140–S148.
13. Rinner,O. and Morgenstern,B. (2002) AGenDA: gene prediction by comparative sequence analysis. *In Silico Biol*., 195–205.
14. Stanke,M., Steinkamp,R., Waack,S. and Morgenstern,B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res*., **32**, W309–W312.
15. Stanke,M. and Morgenstern,B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res*., **33**, W465–W467.
16. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet*., **30**, 13–19.
17. Cawley,S.L. and Pachter,L. (2003) HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics*, **19**, ii36–ii41.
18. Viterbi,A. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform. Theory*, **13**, 260–269.
19. Alexandersson,M., Cawley,S. and Pachter,L. (2003) SLAM—cross-species gene finding and alignment with a generalized pair Hidden Markov Model. *Genome Res*., **13**, 496–502.
20. Abril,J.F. and Guigó,R. (2000) gff2ps: visualizing genomic annotations. *Bioinformatics*, **16**, 743–744.
21. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*., **268**, 78–94.
22. Guigó,R., Knudsen,S., Drake,N. and Smith,T. (1992) Prediction of gene structure. *J. Mol. Biol*., **226**, 141–157.
23. Majoros,W.H., Pertea,M. and Salzberg,S.L. (2004) TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.
24. Gadani,F., Hayes,A., Opperman,C.H., Lommel,S.A., Sosinski,B.R., Burke,M., Hi,L., Brierly,R., Salstead,A. and Heer,J. (2003) Large scale genome sequencing and analysis of *Nicotiana tabacum*: the tobacco genome initiative. *5èmes Journées Scientifiques du Tabac de Bergerac*, pp. 117–130.